PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30

A2

(11) International Publication Number: WO 99/64965

(43) International Publication Date: 16 December 1999 (16.12.99)

(21) International Application Number: PCT/FI99/00492

(22) International Filing Date: 8 June 1999 (08.06.99)

(30) Priority Data:
981355 11 June 1998 (11.06.98) FI

(71) Applicant *(for all designated States except US)*: NOKIA MOBILE PHONES LTD. [FI/FI]; Keilalahdentie 4, FIN–02150 Espoo (FI).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: NIEMI, Terho, Reima, Eerik [FI/FI]; Venuksenkuja 5 F 57, FIN–01480 Vantaa (FI).

(74) Agent: JOHANSSON, Folke; Nokia Corporation, P.O. Box 226, FIN–00045 Nokia Group (FI).

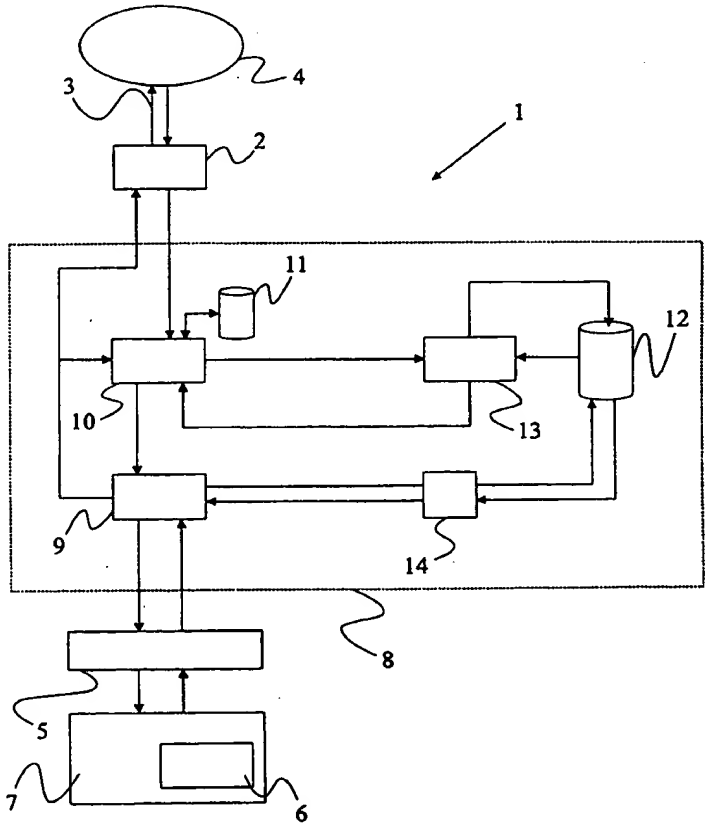
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: ELECTRONIC FILE RETRIEVAL METHOD AND SYSTEM

(57) Abstract

A method of operating a computer system (1) having a display (6) and being connected to the WWW (4). A Web page is downloaded into a buffer memory (8) of the computer system (1) from the WWW (4), the page being an HTML file. Keywords are identified in the downloaded page and the HTML file held in the buffer memory (8) is then modified on–the–fly to introduce links to a database searching application, each link corresponding to an identified keyword. The downloaded Web page is displayed on the computer system display (6), with the introduced links appearing as user selectable items. By selecting a link, the searching application is launched and previously downloaded Web pages which contain the associated keyword are identified.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

ELECTRONIC FILE RETRIEVAL METHOD AND SYSTEM

Field of the Invention

The present invention relates to an electronic file retrieval method and system and more particularly to a method and system which allows a computer system user to retrieve previously processed files which are relevant to a file which is currently being considered by the user.

Background to the Invention

Given the recent dramatic increases in the computer memory available to users of Personal Computers (PCs), users now have the possibility to create massive personal document archives. Added to this is the possibility of retrieving documents over the World Wide Web (WWW) which provides an almost unlimited source of information. Whilst these developments greatly increase the knowledge available to a PC user, it is not often easy for the user to locate information which is relevant to a current task.

Sophisticated search engines have been developed to enable WWW users to "surf" the WWW. These engines, for example AltaVista™, generally operate by exhaustively extracting words from Web pages published on the WWW. Links to these pages are then added to a database alongside the corresponding word entries. Search algorithms have also been developed for searching documents stored, for example, on the hard disc drive of a PC. Again these tend to conduct an exhaustive search of the stored documents for a user defined keyword.

Summary of the Present Invention

It is an object of the present invention to provide a method and system which is able to identify electronic documents or files which are relevant to another electronic document downloaded from a data network, and to incorporate direct or indirect links to those relevant documents or files into the downloaded document.

According to a first aspect of the present invention there is provided a method of operating a computer system, the computer system being connected to a data network and comprising a display and a database storing a set of documents and/or document identifiers, the method comprising:

downloading an electronic document into the computer system over the data network, the document being in the form of computer readable code;

identifying keywords in the downloaded document;

modifying said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored or identified in said database and containing at least one of said keywords; and

displaying the downloaded document on the computer system display, where the introduced hyper-links appear as user selectable items.

Preferably, said step of identifying keywords comprises defining a global keyword list on the basis of the stored and/or identified documents and the downloaded document, and identifying those of the global keywords present in the downloaded document. More preferably, the global keyword list is defined by analysing word occurrence rate distribution over the documents.

Preferably, the said step of modifying the computer readable code comprises creating a hyper-link for each identified keyword. More preferably, the method comprises the further steps of:

activating one of said introduced hyper-links;

determining which of the stored and/or identified documents contain the associated keyword;

determining a similarity/dissimilarity coefficient for each of these documents; and

displaying a list of these documents together with the similarity/dissimilarity coefficient.

The displayed list preferably includes hyper-links to the listed documents.

Preferably, hyper-links are displayed by highlighting, e.g. using a colour change, underlining, or italicising, the keywords identified in the downloaded document. Links may also be displayed as specific characters, e.g. ".", "!", "?".

Preferably, said computer readable code is Hyper Text Markup Language (HTML), in which case said steps of downloading and displaying may be performed by a Web browser.

Preferably, the data network over which the electronic document is downloaded is the World Wide Web. The step of displaying the downloaded document (with added links) may involve interpreting the document with an Internet Browser.

The identifier may be a document title, a computer drive path, or Universal Resource Locator (URL) to an Web page, or a combination of these. The documents used to construct the database may include Web pages, word processed documents, and electronic mail items.

According to a second aspect of the present invention there is provided a programmed computer system comprising;

communication means coupled to a data network for downloading an electronic document over the data network, the document being in the form of a computer readable code;

an electronic database storing a set of documents and/or document identifiers;

first processing means arranged in use to identify keywords in said downloaded document;

second processing means arranged in use to modify said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored or identified in said database and containing at least one of said keywords; and

a display and display driver means arranged in use to display the downloaded document on the computer system display in a form where the introduced hyper-links appear as user selectable items.

In certain embodiments of the above second aspect of the present invention, the computer system is provided by a suitably programmed computer, where the communication means is a data modem of the computer, and the first and second processing means comprises a microprocessor or a digital signal processor.

In other embodiments of the invention the system comprises a personal computer connected to a local area network, which is coupled to the WWW via a router. Said database and said first and second processing means may be provided in said personal computer or in a second computer also connected to the local area network and accessible by other personal computers. Alternatively, the database and first and second processing means may be replicated in the personal computer and in one or more other computers of the local area network, to provide a hierarchy of "knowledge" servers.

According to a third aspect of the present invention there is provided a computer memory encoded with executable instructions representing a computer program for causing a computer system connected to a data network to:

construct an electronic database storing a set of documents and/or document identifiers;

download an electronic document into the computer system over the data network, the document being in the form of a computer readable code;

identify keywords in said document;

modify said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored or identified in said database and containing at least one of said keywords; and

display the downloaded document on the computer system display, where the introduced hyper-links appear as user selectable items.

Brief Description of the Drawings

For a better understanding of the present invention and in order to show how the same may be carried into effect reference will now be made, by way of example, to the accompanying drawings, in which:

Figure 1 shows schematically a computer connected to the WWW and arranged to construct and make use of a knowledge database; and

Figure 2 shows a displayed Web page with hyper-links added on-the-fly.

Detailed Description of Certain Embodiments

With reference to Figure 1, a personal computer (PC) 1 is shown functionally and is indicated generally by the reference numeral 1. The PC 1 has a modem 2 which enables the PC 1 to be connected to a telephone line 3

and via the telephone line to the WWW 4. It will be understood that this connection may additionally involve an Internet Service Provider (ISP) although this is not shown in the Figure. It will also be appreciated that the PC 1 may alternatively be connected to the WWW 4 via a local area network (LAN) having its own Network Access Server (NAS).

Using a computer program stored in a memory of the PC 1 and executed by a central processing unit, the PC 1 is provided with an Internet browser 5. This may be a conventional browser such as Microsoft Internet Explorer™ or Netscape Navigator™, but in any case is capable of requesting and retrieving a Web page from the WWW 4 via the modem 2 using http or another suitable Internet protocol and a Universal Resource Locator (URL) or similar identifier (e.g. Universal Resource Name) identifying the page.

As is already well known, the WWW makes use of a special programming language or code known as Hyper Text Mark-up Language (HTML) to encode Web pages, and the Internet browser 5 is capable of interpreting this code and causing a received page to be displayed on a display 6 of a user interface 7 of the PC 1. Using HTML, so called "hyper-links" may be incorporated into a Web page. Hyper-links are elements of a Web page, e.g. legends, word, pictures etc, which the user may select using a computer mouse. Selecting a hyper-link normally causes the browser 5 to download a further Web page from the WWW 4, which page is identified in the originally received HTML by a URL associated with that particular hyper-link.

Figure 1 illustrates an additional software/hardware module 8 which is functionally placed between the Web browser 5 and the modem 2. The module may be implemented

by suitably programming the CPU of the PC 1 or using a DSP or the like. The module 8 comprises a server 9 which communicates with the Web browser using TCP/IP. In the event that a user requests the downloading of a Web page from the WWW 4 by entering a URL, this request is relayed by the server 9 (acting as a proxy server) to the WWW. In relaying the request, the proxy server 9 also passes the URL to a memory control function 10 which stores the URL in an associated memory block 11.

When the requested Web page is returned from the WWW 4, the page is intercepted by the module 8 (on the basis of the stored URL) and is temporarily stored in the memory block 11 by the memory control function 10. Before forwarding the page to the Internet browser 5 for display, the page is modified in the memory block 11 as will be described below. Firstly however, it is necessary to explain the structure and function of a keyword database which is stored in a memory block identified by reference numeral 12 in Figure 1.

The database 12 contains a "word" table which lists in a first column every word stem which appears in at least one previously analysed document (in this example previously downloaded Web pages). For example, the words "produce", "produced", "produces", "producing", etc are represented in the table by a single stem "produc" (a suitable "stemming" algorithm is described in "Development of a Stemming Algorithm", Julie Beth Lovins, Mechanical Translation and Computational Linguistics, 11, 22-31, 1968). For each word (stem), the number of documents in which there are 0 occurrences of the word is entered in column 2, the number of documents in which there is 1 occurrence of the word is entered in column 3, etc. This is illustrated by Table 1 below from which it can be seen

that word #3 occurs 0 times in 45 documents, 1 time in 1 document, etc.

Using the information contained in the above table, it is possible to determine which of the words listed are likely to be keywords. This process is described in two articles titled "A Probabilistic Approach to Automatic Keyword Indexing", Part I, Journal of the American Society for Information Science, July-August 1975, pp 197-206; Part II, Journal of the American Society for Information Science, September-October 1975, pp 280-289. Briefly, the process involves determining, for each listed word, the Poisson distribution of the rate of occurrence and determining the error between the actual distribution and the Poisson distribution. A constant is then defined, and words for which the error is less than that constant are identified as keywords. Those listed words which are identified as being keywords are identified in an additional column of the above table by a keyword flag, set to 1 if the word is a keyword and 0 if it is not a keyword. These keywords are global in the sense that they are derived on the basis of all analysed documents.

The database 12 contains a second table referred to as the "document" table. This table contains for each examined document the number of occurrences of each word together with the URL identifying the document. An example of such a table is shown below in Table 2 from which it can be seen that document 2 contains a single occurrence of word #1, 23 occurrences of word #2 etc.

It will be appreciated that when the system is first initiated, the word and document tables may be empty. These tables are then built-up as the system is used. It

is of course also possible to pre-install tables based on the analysis of a number of representative documents.

Returning now to the downloaded Web page temporarily stored in the memory block 11, it will be appreciated that this page is typically in HTML format and is likely to contain several portions of text. This text is extracted by a text analysis function 13, and the word table stored in the database 12 is updated on the basis of the extracted text. The document table is similarly updated. On the basis of the updated word table, the keyword list is refined (and the keyword flags set accordingly).

The text analysis function 13 then scans the text contained in the downloaded Web page to identify keywords present therein. When a keyword is identified, the function 13 modifies the HTML code held in the buffer "on-the-fly", to introduce an associated hyper-link (the function of which is explained below). The following HTML listing shows a downloaded Web page which contains the keywords "TeamWARE"™, "Internet", "desktop", and "agents", and in which added hyper-links are shown underlined (this page did not contain any original hyper-links).

```
<HTML>
<HEAD>
<TITLE>TeamWARE Mail 5</TITLE>
</HEAD>

<BODY bgcolor="ffffff" TEXT="#000000" link="#ff0000" vlink="#990000" alink="#ff0000">

<blockquote>

<H1><A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top"
STYLE="color: green">TeamWARE</A> Mail 5</H1>
<H2>For Business Critical Messaging</h2>

<P> </p>
```

<h3><A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail 5 is a ready-to-run</h3>

<B>e-mail service combining the ease-of-use
and global reach of <A HREF="http://niemi_terho/page-62/word-34488/default" TARGET="_top" STYLE="color: green">Internet</A> SMTP/MIME e-mail, access to LDAP/X.500
based directories, with the security, reliability, and maintainability
of an Intranet messaging system. <A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail 5 offers X.400
connectivity to reach those important business partners who prefer
X.400-based mail. The Connector for Fax enables sending faxes
in the same easy way as sending e-mail. Mobile users can utilize
the Connector for SMS to receive notifications through GSM network.</B>

<P>
<B><A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail 5 is modular and scaleable,
which allows you to add functionality when you need it, and secures
the growth path for your messaging system. <A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail's key features are:</B><p>
<A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail 5 Server:
<UL>
<LI>Advanced <A HREF="http://niemi_terho/page-62/word-34488/default" TARGET="_top" STYLE="color: green">Internet</A> Mail (SMTP/MIME), X.400(88)
Mail and Connectors for Fax and SMS
<LI>Message and attachment compression, mailbox and
message size control
<LI>Windows-based user and system administration,
On-line backup for 7 x 24 operations, Incremental backup support
<LI>Customisable Organization and Personal Directories,
transparent LDAP search and directory synchronization
<LI>Support for IMAP4 and POP3 client protocols
</UL>

<h3><A HREF="http://niemi_terho/page-62/word-29329/default" TARGET="_top" STYLE="color: green">TeamWARE</A> Mail 5 Client:</h3>

<UL>
<LI>Rich Text Format (RTF) mail editor, Reminders,
Server-based notifications
<LI>Searchable, hierarchical mail folders
<LI>Mailbox permissions, private folders
<LI><A HREF="http://niemi_terho/page-62/word-34490/default" TARGET="_top" STYLE="color: green">Desktop</A> integration through Simple MAPI
<LI><A HREF="http://niemi_terho/page-62/word-523/default" TARGET="_top" STYLE="color: green">Agents</A> for executing previously defined tasks
on behalf of users
</UL>

</BODY>
</HTML>

After modification of the HTML code on-the-fly, the Web page is returned from the memory block 11 to the Web browser 5 via the memory control function 10 and the server 9. Figure 2 shows the Web page corresponding to the above code as displayed by the Web browser 5. The added hyper-links are shown underlined, although it will be appreciated (from the above code) that when displayed in colour these will appear green.

Assume now that the user wishes to locate documents which are related to the downloaded page, and in particular are related by way of the keyword "Internet". By clicking on one of the "Internet" hyper-links added to the Web page, the user causes the Web browser 5 to request from the server 9 (acting as Web server) the contents of the URL "http://niemi_terho/page-62/word-29329/default". The "default" identifier contained in this URL causes the Web server 9 to launch an application 14 termed a Dynamic Linkable Library (DLL) which links the Web server 9 to the database 12.

The URL word identifier, in this case "word-29329", identifies to the database 12 the keyword "Internet" in the word table, whilst the page identifier, "page-62", identifies the source document, i.e. the downloaded Web page, in the document table. The database 12 first of all identifies all of the documents identified therein which contain the keyword "Internet". A dissimilarity coefficient is then calculated for each of the identified documents, relative to the source document, as follows.

Consider that eight documents (1 to 8) containing the keyword "Internet" are identified, and that each of these contains one or more of the complete set of listed keywords (I to P, where I represents the keyword

"Internet"). If "Ref" is the downloaded document (i.e. page-62), then this information can be represented graphically as illustrated in Table 3 below.

The dissimilarity coefficient for an identified document is calculated as follows:

$$\frac{|X \Delta Y|}{|X| + |Y|}$$

where X is the set of keywords contained in document Ref, Y is the set of keywords contained in the identified document in question, and $X \Delta Y = (X \cup Y) - (X \cap Y)$. So, for example, in the case of document 1, the numerator in the above equation equals 3 (i.e. 6 shared keywords minus 3 not shared keywords) whilst the denominator equals 9, giving a dissimilarity coefficient of 0.333 or 33.3%.

The DLL 14 returns from the database 12, to the Web server 9 and the Web browser 5, the URLs of the identified pages together with the respective dissimilarity coefficients. These are displayed to the user as a list of hyper-links. Assuming that the user selects one such hyper-link, the Web browser 5 causes the selected Web page to be downloaded over the WWW 4 and displayed on the display 6.

In order to prevent common words, such as "the", "and", from being identified as keywords, a STOP list may be defined which contains words which cannot be included in the word table. However, in order to provide for increased flexibility, it is preferred to include in the word table a keyword "lifetime" value which represents a number of days.

If the lifetime is negative, a word cannot become a keyword until the lifetime value reaches 0. So, for example, the lifetime of the word "and" may be set to -1000,000 days so that effectively it can never become a keyword. For a word such as "computer", which tends to be somewhat more distinctive, the lifetime may be set to -30 days. It can be expected that after a period of 30 days, the system will have been "educated" sufficiently to determine whether or not the term "computer" is indeed a relevant keyword.

On the other hand, the lifetime may be positive so that a user may force a word onto the keyword list for some fixed period of time. For example, if a user is currently interested in topics related to the Internet, then the lifetime of the word "Internet" could be defined as +30 days. Thus, "Internet" would be retained as a keyword for 30 days, after which its retention depends upon the keyword determination process described above.

The process described above may be easily modified to enable a user to identify Web pages which are associated to a downloaded Web page by something other than conventional keywords. For example, association may be carried out on the basis of names, dates, or tasks, or on the basis of some combination of these.

The process may also be extended to provide a hierarchy of "knowledge servers" forming a chain between the WWW 4 and the end user interface 7, each such server being provided with a module 8. Servers higher up in the chain will tend to serve a number of different PCs 1. For example, there may be provided a corporate knowledge server and a group knowledge server between the end user PC 1 and the WWW, in which case the corporate server collects keywords from Web pages downloaded to all users

in the company, whilst the group server collects keywords from pages downloaded to only the members of a particular group.

It is also possible to extend the process to incorporate into the database details of non-Web page electronic files. For example, keywords could be identified in word processor generated files or in electronic mail received or sent by the PC 1. The location (i.e. drive paths) of these files could then be stored as the file links in the database 10. In the case of electronic mail, the function of the Web browser 5 may be replaced by an e-mail "client" such as Microsoft Outlook Express™. Typically the client then connects via the module 8 to a mail server.

The person of skill in the art will appreciate that modifications may be made to the above described embodiments without departing from the scope of the present invention. In particular, in computing the dissimilarity coefficient between a downloaded document and a previously archived document, account may be taken of the number of occurrences of a keyword in the documents. For example, Table 3 may be modified to replace the "✔" signs with the number of occurrences of a given keyword in a given document. The equation described above for calculating the dissimilarity coefficient is replaced by a new equation in which the numerator is the sum of the absolute difference for each keyword, whilst the numerator is the total number of keywords in the two compared documents.

Considering for example Table 3, assume that the keywords I, J, K, and L appear 2, 3, 1, and 4 times respectively in document 1, whilst the keywords I, J, L, M, and O appear 1, 3, 2, 4, and 2 times respectively in document

Ref. The numerator of the dissimilarity coefficient is given by:

$$|2-1| + |3-3| + |1-0| + |4-2| + |0-4| + |0-2| = 10$$

whilst the denominator is given by:

$$(2+3+1+4) + (1+3+2+4+2) = 22$$

so that the dissimilarity coefficient is 45.5%.

The calculation of the dissimilarity coefficient may be further modified to take into account document length. For example, the number of occurrences of a keyword in a document may be normalised by a normalising factor equal to the length of standard document divided by the length of the document in question.

	Number of occurrences/document							
Word	0	1	2	3	4	5	etc	keyword flag
#1	8	1	15	21	1	7	—	0 or 1
#2	6	45	23	6	26	17	—	0 or 1
#3	45	1	34	10	18	12	—	0 or 1
#4	23	4	56	34	29	28	—	0 or 1
etc	—	—	—	—	—	—	—	0 or 1

Table 1

Document	Word #1	Word #2	Word #3	etc	URL
1	0	0	3	—	www. etc
2	1	23	5	—	www. etc
3	0	12	5	—	www. etc
etc	—	—	—	—	www. etc

Table 2

	I	J	K	L	M	N	O	P
1	✔	✔	✔	✔				
2	✔	✔	✔					
3	✔			✔				
4	✔		✔	✔		✔		
5	✔				✔	✔	✔	
6	✔				✔	✔	✔	
7	✔				✔	✔		✔
8	✔				✔	✔		✔
Ref	✔	✔		✔	✔		✔	

Table 3

Claims

1. A method of operating a computer system, the computer system being connected to a data network and comprising a display and a database storing a set of documents and/or document identifiers, the method comprising:

downloading an electronic document into the computer system over the data network, the document being in the form of computer readable code;

identifying keywords in the downloaded document;

modifying said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored or identified in said database and containing at least one of said keywords; and

displaying the downloaded document on the computer system display, where the introduced hyper-links appear as user selectable items.

2. A method according to claim 1, wherein said step of identifying keywords comprises defining a global keyword list on the basis of the stored and/or identified documents and the downloaded document, and identifying those of the global keywords present in the downloaded document.

3. A method according to claim 2, wherein the global keyword list is defined by analysing word occurrence rate distribution over the documents.

4. A method according to claim 1, wherein said step of modifying the computer readable code comprises creating a hyper-link for each identified keyword.

5. A method according to claim 4, wherein the method comprises the further steps of:

activating one of said introduced hyper-links;

determining which of the stored and/or identified documents contain the associated keyword;

determining a similarity/dissimilarity coefficient for each of these documents; and

displaying a list of these documents together with the similarity/dissimilarity coefficient.

6. A method according to claim 5 and comprising displaying said list as a set of hyper-links to the listed documents.

7. A method according to claim 1, wherein said computer readable code is Hyper Text Markup Language (HTML), and said steps of downloading and displaying are performed by a Web browser.

8. A method according to claim 1, wherein the data network over which the electronic document is downloaded is the World Wide Web.

9. A method according to claim 1, wherein the document identifiers are a document title, a computer drive path, or Universal Resource Locator (URL) to an Web page, or a combination of these.

10. A programmed computer system comprising;

communication means (5,2) coupled to a data network (4) for downloading an electronic document over the data network (4), the document being in the form of a computer readable code;

an electronic database (12) storing a set of documents and/or document identifiers;

first processing means (12,13) arranged in use to identify keywords in said downloaded document;

second processing means (13) arranged in use to modify said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored

or identified in said database (12) and containing at least one of said keywords; and

a display (6) and display driver means (5,7) arranged in use to display the downloaded document on the computer system display in a form where the introduced hyper-links appear as user selectable items.

11. A system according to claim 10, wherein the computer system is provided by a suitably programmed computer (1), where the communication means (2,5) comprises a data modem of the computer (1), and the first and second processing means (12,13) comprise a microprocessor or a digital signal processor.

12. A computer memory encoded with executable instructions representing a computer program for causing a computer system connected to a data network to:

construct an electronic database storing a set of documents and/or document identifiers;

download an electronic document into the computer system over the data network, the document being in the form of a computer readable code;

identify keywords in said document;

modify said computer readable code to introduce thereinto hyper-links to enable a user to link to documents stored or identified in said database and containing at least one of said keywords; and

display the downloaded document on the computer system display, where the introduced hyper-links appear as user selectable items.

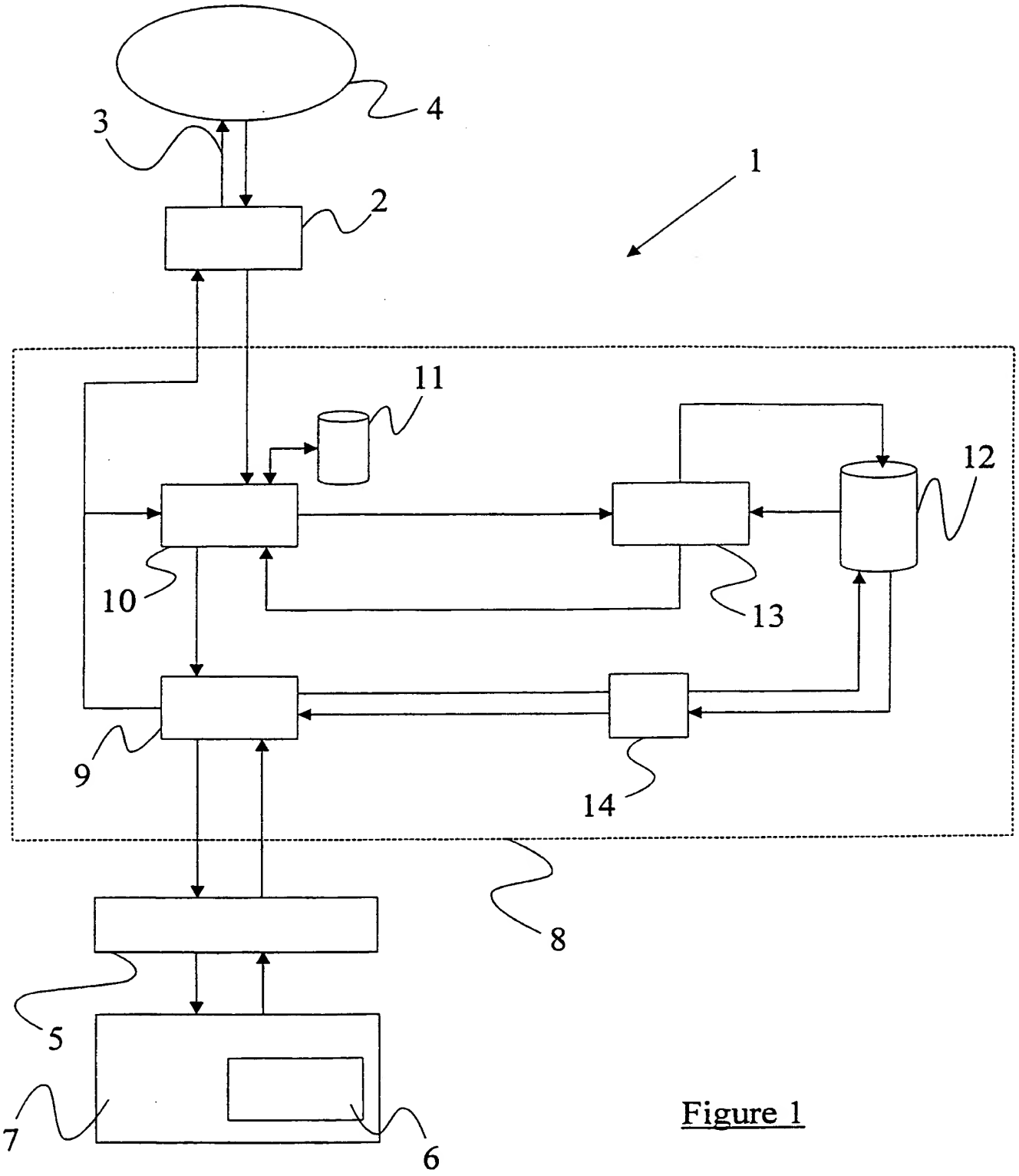


Figure 1

TeamWARE Mail 5

For Business Critical Messaging

TeamWARE Mail 5 is a ready-to-run

e-mail service combining the ease-of-use and global reach of Internet SMTP/MIME e-mail, access to LDAP/X.500 based directories, with the security, reliability, and maintainability of an Intranet messaging system. TeamWARE Mail 5 offers X.400 connectivity to reach those important business partners who prefer X.400-based mail. The Connector for Fax enables sending faxes in the same easy way as sending e-mail. Mobile users can utilize the Connector for SMS to receive notifications through GSM network.

TeamWARE Mail 5 is modular and scaleable, which allows you to add functionality when you need it, and secures the growth path for your messaging system. TeamWARE Mail's key features are:

TeamWARE Mail 5 Server:

- Advanced Internet Mail (SMTP/MIME), X.400(88) Mail and Connectors for Fax and SMS •Message and attachment compression, mailbox and message size control
- Windows-based user and system administration, On-line backup for 7 x 24 operations, Incremental backup support
- Customisable Organization and Personal Directories, transparent LDAP search and directory synchronization
- Support for IMAP4 and POP3 client protocols

TeamWARE Mail 5 Client:

- Rich Text Format (RTF) mail editor, Reminders, Server-based notifications
- Searchable, hierarchical mail folders
- Mailbox permissions, private folders
- Desktop integration through Simple MAPI
- Agents for executing previously defined tasks on behalf of user

Figure 2

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/30

A3

(11) International Publication Number: **WO 99/64965**

(43) International Publication Date: 16 December 1999 (16.12.99)

(21) International Application Number: PCT/FI99/00492

(22) International Filing Date: 8 June 1999 (08.06.99)

(30) Priority Data:
981355 11 June 1998 (11.06.98) FI

(71) Applicant *(for all designated States except US)*: NOKIA MOBILE PHONES LTD. [FI/FI]; Keilalahdentie 4, FIN–02150 Espoo (FI).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: NIEMI, Terho, Reima, Eerik [FI/FI]; Venuksenkuja 5 F 57, FIN–01480 Vantaa (FI).

(74) Agent: JOHANSSON, Folke; Nokia Corporation, P.O. Box 226, FIN–00045 Nokia Group (FI).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

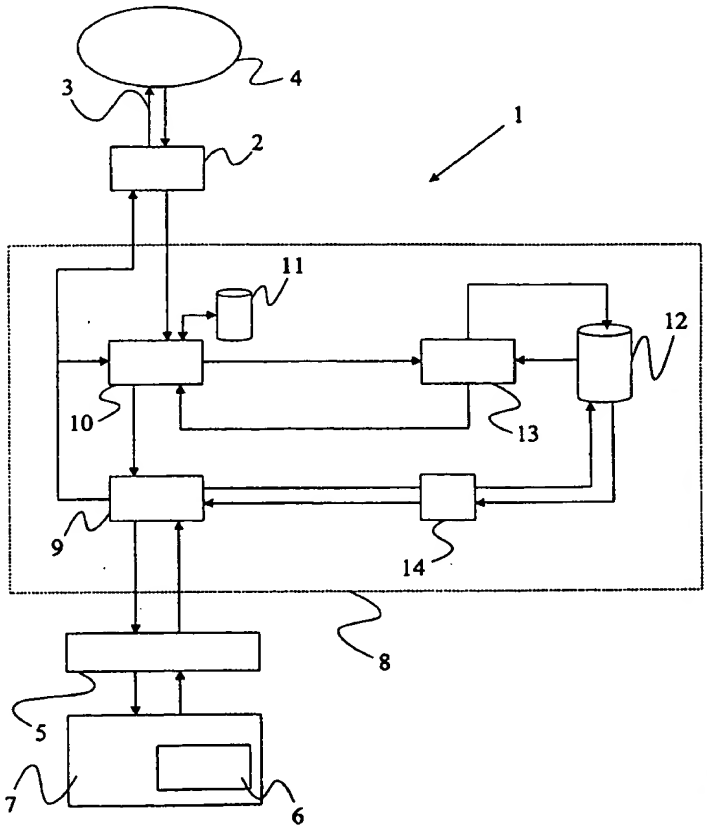
Published
With international search report.

(88) Date of publication of the international search report:
17 February 2000 (17.02.00)

(54) Title: ELECTRONIC FILE RETRIEVAL METHOD AND SYSTEM

(57) Abstract

A method of operating a computer system (1) having a display (6) and being connected to the WWW (4). A Web page is downloaded into a buffer memory (8) of the computer system (1) from the WWW (4), the page being an HTML file. Keywords are identified in the downloaded page and the HTML file held in the buffer memory (8) is then modified on–the–fly to introduce links to a database searching application, each link corresponding to an identified keyword. The downloaded Web page is displayed on the computer system display (6), with the introduced links appearing as user selectable items. By selecting a link, the searching application is launched and previously downloaded Web pages which contain the associated keyword are identified.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI 99/00492

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G06F 17/30
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0810534 A2 (OKI ELECTRIC INDUSTRY CO., LTD.), 3 December 1997 (03.12.97), column 2, line 57 - column 3, line 11; column 39, line 34 - column 40, line 46, abstract --	1-12
X	EP 0778534 A1 (SUN MICROSYSTEMS, INC.), 11 June 1997 (11.06.97), page 4, line 27 - line 59, abstract --	1-12
P,A	GB 2329988 A (INTERNATIONAL BUSINESS MACHINES CORPORATION), 7 April 1999 (07.04.99), page 2, column 32 - page 3, column 36, abstract --	1-12

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
6 December 1999	09 -12- 1999
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Joni Sayeler/AE Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI 99/00492

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E,A	EP 0926606 A2 (CASIO COMPUTER CO., LTD.), 30 June 1999 (30.06.99), see the whole document -- --------	1-12

INTERNATIONAL SEARCH REPORT

Information on patent family members

02/11/99

International application No.
PCT/FI 99/00492

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0810534 A2	03/12/97	CA 2204447 A	13/11/97
		JP 9305579 A	28/11/97
		JP 9305623 A	28/11/97
		JP 9305475 A	28/11/97
		JP 9305581 A	28/11/97
EP 0778534 A1	11/06/97	AU 706512 B	17/06/99
		AU 7185596 A	12/06/97
		CA 2191671 A	09/06/97
		CN 1157965 A	27/08/97
		JP 10049425 A	20/02/98
		US 5822539 A	13/10/98
GB 2329988 A	07/04/99	GB 9818505 D	00/00/00
EP 0926606 A2	30/06/99	JP 11195025 A	21/07/99